week11-scholar

Mathi Manavalan

4/1/2020

R Studio API Code

Setting the working directory is not necessary for an RMarkdown file.

Libraries

Importing the libraries necessary for the data importing, cleaning, and analyzing.

```
library(tidyverse)
library(rvest)
```

Data Import and Cleaning

Here, I am web-scraping some data from Dr. Yuhong Jiang's Google Scholar page.

```
papers <- read_html("https://scholar.google.com/citations?user=ZqlRcM8AAAAJ&hl=en&oi=ao")

titleNodes <- html_nodes(papers, ".gsc_a_at")
titleText <- html_text(titleNodes)

authorNodes <- html_nodes(papers, ".gsc_a_at+ .gs_gray")
authorText <- html_text(authorNodes)

yearNodes <- html_nodes(papers, ".gsc_a_hc")
yearText <- html_text(yearNodes)

countNodes <- html_nodes(papers, ".gsc_a_ac")
countText <- html_text(countNodes)

profile_tbl <- cbind(titleText, authorText, yearText, countText) %>%
  as_tibble() %>%
  mutate(countText = as.numeric(countText), yearText = as.numeric(yearText))
```

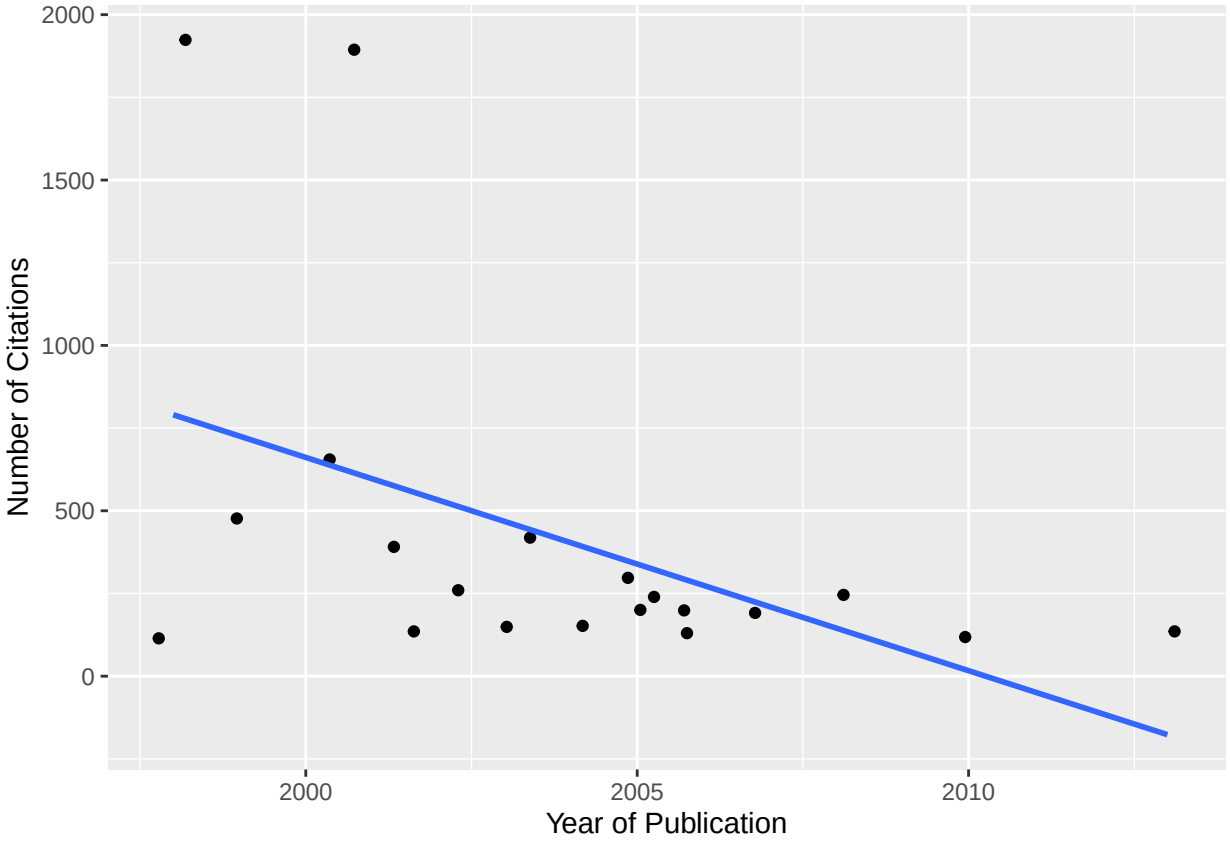
Analysis

```
sum <- summary(lm(profile_tbl$countText ~ profile_tbl$yearText))
```

The correlation between year and count is -64.4809004.

Visualization

```
plot <- ggplot(profile_tbl, aes(x=yearText, y= countText)) +
  geom_jitter() +
  labs(x = "Year of Publication", y = "Number of Citations") +
  geom_smooth(method = "lm", se = FALSE)
plot
```



The above plot displays the relationship between the year of publication and the number of citations of some of Dr. Yuhong Jiang's publications. The plot also includes a linear regression line, modeling the relationship between the two variables linearly.